PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

CrossMark click for updates

Phil. Trans. R. Soc. B (2012) **367**, 2574–2583
doi:10.1098/rstb.2012.0010

Research

Protein kinases display minimal interpositional dependence on substrate sequence: potential implications for the evolution of signalling networks

Brian A. Joughin[1], Chengcheng Liu[3], Douglas A. Lauffenburger[1,2], Christopher W. V. Hogue[3] and Michael B. Yaffe[1,2,*]

[1]*The David H. Koch Institute for Integrative Cancer Research, and* [2]*Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*
[3]*Computation and Systems Biology Programme, Singapore-MIT Alliance, National University of Singapore, Singapore, Republic of Singapore*

Characterization of *in vitro* substrates of protein kinases by peptide library screening provides a wealth of information on the substrate specificity of kinases for amino acids at particular positions relative to the site of phosphorylation, but provides no information concerning interdependence among positions. High-throughput techniques have recently made it feasible to identify large numbers of *in vivo* kinase substrates. We used data from experiments on the kinases ATM/ATR and CDK1, and curated CK2 substrates to evaluate the prevalence of interactions between substrate positions within a motif and the utility of these interactions in predicting kinase substrates. Among these data, evidence of interpositional sequence dependencies is strikingly rare, and what dependency exists does little to aid in the prediction of novel kinase substrates. Significant increases in the ability of models to predict kinase–substrate specificity beyond position-independent models must come largely from inclusion of elements of biological and cellular context, rather than further analysis of substrate sequences alone. Our results suggest that, evolutionarily, kinase substrate fitness exists in a smooth energetic landscape. Taken with results from others indicating that phosphopeptide-binding domains do exhibit interpositional dependence, our data suggest that incorporation of new substrate molecules into phospho-signalling networks may be rate-limited by the evolution of suitability for binding by phosphopeptide-binding domains.

Keywords: phosphorylation; kinase substrate specificity; signalling networks

1. INTRODUCTION

Substrate recognition by kinases is mediated both by recognition of the substrate sequence surrounding the phosphorylation site and by other contextual elements including co-localization and distal-site interaction. The specificity of protein kinases for their substrate sites has traditionally been described at one of three levels of complexity. The first level is that of the match to a phosphorylation motif [1], a description in terms of which amino acids are frequently present at which positions relative to the site of phosphorylation (e.g. 'S/T-P-X-R/K', for the cell-cycle-dependent kinase Cdk1 [2]). These descriptions are practical and readable, but are generally neither sufficient nor necessary descriptions of a kinase's substrates [2]. At a more complex level, the *in vitro* specificity of a kinase can be described using a position-specific scoring matrix (PSSM), wherein the probability of each amino acid at each position relative to the site of phosphorylation is given. Any higher-order information on what amino acids are more or less likely to co-occur in the positions surrounding the site of the phosphorylation is lost [3]. As a result, the suitability of non-ideal kinase substrates can be captured, instead of only substrates that match the most likely amino acid at each position. Both of these descriptions represent simple models of the specificity of a kinase for its substrates, using only sequence information local to the potential site of phosphorylation. The motif description is a simple Boolean descriptor: to an extreme approximation, the kinase will phosphorylate sites that fit the motif, and will not phosphorylate those that do not. The PSSM description is more nuanced: given a sequence, a score can be calculated that should approximately reflect the suitability of the substrate sequence for the kinase. At the final, most complex level, several models of kinase–substrate specificity more complex than the PSSM have been proposed. These include neural networks (NetPhosK [4] and NetPhorest [5]), hidden Markov models

* Author for correspondence (myaffe@mit.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2012.0010 or via http://rstb.royalsocietypublishing.org.

One contribution of 13 to a Theme Issue 'The evolution of protein phosphorylation'.

(KinasePhos) [6], support vector machines (PredPhospho [7]) and Bayesian decision theory (PPSP [8]). These more detailed models are based on the assumption that better predictions of kinase substrates can be obtained if machine-learning algorithms are allowed to use information of higher-order in developing classifiers. One potential failing of these methods is the fact that they use large databases of substrate–kinase pairs identified by different experimental methods under different conditions, such as Phospho.ELM [9] and UniProtKB/Swiss-Prot [10]. While this provides a large training set, it also poses the risk of convolving kinase specificity with additional biases present owing to different portions of the source data being collected by different methods.

It is important to determine whether higher-order models of amino acid sequence specificity than that provided by a PSSM can improve description of the *in vivo* activity of a kinase or, more especially, improve prediction of kinase–substrate interactions, rather than overfit data or fit co-dependencies created by study biases instead of the kinase's own biophysical selectivity. If a kinase does not recognize each amino acid position on its substrate in an energetically independent manner, then models for describing and predicting kinase substrates should take this into account. If, on the other hand, the assumptions of energetic independence that lead to models such as the PSSM hold true, then future improvements in our ability to predict novel kinase substrates will depend instead on identifying and quantifying the contribution of other factors important to the kinase–substrate interaction, such as recognition via distal interaction sites, protein localization and dynamics and the presence or absence of other facilitating or inhibitory post-translational modifications [11].

Of note, the majority of protein kinases do not have a known motif specificity; a recent atlas of computationally predicted consensus specificities covers only 35 per cent of the human kinome [5]. Conversely, only 12 per cent of the curated phosphorylation sites in Phospho.ELM are currently attributable to a particular kinase [9]. For most kinases, a large set of known substrates gathered by a single method does not exist. Nonetheless, there are a small number of single experiments described in the literature that identify a large number of kinase substrates. From these datasets, it may be possible to gain an understanding of the role of cooperativity among substrate amino acid positions in determining kinase specificity. We therefore set out to assess directly the degree to which statistically significant positional interdependencies among kinase substrate sequences can be identified. For this work, three recently available datasets including substrates of three highly pleiotropic kinases were selected. Matsuoka *et al.* [12] identified 905 putative substrate sites for the DNA-damage-dependent kinases ATM (ataxia telangiectasia mutated) and ATR (ataxia telangiectasia and Rad3 related) (and possibly other ionizing radiation-induced kinases such as DNA-PK) using quantitative mass spectroscopy. Blethrow *et al.* [13] identified 71 direct proline-directed human substrate sites for Cdk1/Cyclin B, using a kinase modified to accept a chemically labelled adenosine triphosphate (ATP) analogue, followed by covalent capture and identification of labelled sites via mass spectrometry. While there are not as many substrate sites available in this dataset as in the ATM/ATR dataset, these data are extremely appealing, in that the data are direct observations of the substrates of a single kinase in a single experiment; there are few potential sources of external bias. Finally, Salvi *et al.* [14] have curated a set of 492 putative substrate sites, from a number of species, of casein kinase2 (CK2). These sites were not identified in a single experiment; just over 300 were identified from the literature by Meggio & Pinna [15], while the remainder were extracted from the large-scale databases PhosphoSite [16] and Phospho.ELM [9]. While some of the problems of heterogeneous data may be present in this dataset, it is possible that curation by experts in the particular kinase, CK2, may mitigate these problems.

2. RESULTS

(a) *Statistical significance of interpositional dependencies among kinase substrates*

We first interrogated the degree to which second-order substrate preferences can be identified in datasets associated with substrates of the DNA-damage-dependent kinases ATM and ATR [12]. At substrate positions relative to the fixed site of phosphorylation, we counted the number of occurrences of each amino acid. At each pair of positions, we counted the number of co-occurrences of each pair of amino acids. In order to determine the significance of enrichment (or reduction) of co-occurrence of a pair of amino acids, we then compared the actual degree of co-occurrence with the distribution of co-occurrences that might be expected by chance given the individual levels of occurrence for each amino acid (see §4*c*,*d*). We expected to find a large number of significant deviations, driven by kinase structural biophysics or by evolutionary selection for some downstream functional constraint. Contrary to this expectation, although we observed small deviations in the frequencies of individual amino acid pairs from what might be most expected by chance, these were rarely statistically significant when subjected to rigorous analysis (table 1). That is, almost all of these deviations correspond to what would be expected under a model of position-wise independence by random chance.

Despite a large dataset of 861 substrate sites of ATM and ATR, only a handful of statistically significant deviations from what would be expected by chance under a position-independent model can be identified (table 1). Each of these involves the phosphorylated position (i.e. whether a serine or threonine is phosphorylated). Phosphoserine at the fixed position occurs more frequently than can be explained by chance with proline or glycine at the -1 position, with glycine at the $+2$ position and with serine at the $+3$ position. Conversely, phosphothreonine co-occurs less often than can be explained by chance with these amino acids. This result is likely caused by the statistical boosting power of the oriented phosphorylated position; while a pair of arbitrary positions may populate a space of 20×20 amino acid pairs, pairs of positions including the phosphorylated serine or threonine exist in a space of only

Table 1. Substrate sequence position pairs demonstrating significant deviations from position-wise independence.

kinase	position 1	position 2	motif[a]	type	*p*-value[b]
ATM/ATR	0	2	**pS**-Q-G	enriched	4.93×10^{-4}
	0	2	**pT**-Q-G	reduced	4.93×10^{-4}
	0	3	**pS**-Q-X-S	enriched	1.16×10^{-3}
	0	3	**pT**-Q-X-S	reduced	1.16×10^{-3}
	−1	0	Struct.-**pS**-Q	enriched	9.23×10^{-5}
	−1	0	Struct.-**pT**-Q	reduced	9.23×10^{-5}
Cdk1/Cyclin B	3	4	**pS/pT**-P-X-Basic-Polar	reduced	1.63×10^{-3}
CK2	1	2	**pS/pT**-E-E	enriched	2.25×10^{-5}
	2	4	**pS/pT**-X-Polar-X-Polar	enriched	1.39×10^{-3}
	4	5	**pS/pT**-X-X-X-Acidic-Acidic	enriched	2.82×10^{-4}
	2	4	**pS/pT**-X-Hyd-X-Struct.	enriched[c]	3.14×10^{-3}
	2	5	**pS/pT**-X-Hyd.-X-X-Hyd.	enriched[c]	1.29×10^{-3}

[a]Grouped amino acid definitions: structural (G,P), basic (K,R), acidic (D,E), polar (C,H,N,Q,S,T), hydrophobic (A,I,L,M,V), aromatic (F,Y,W). Bold entries indicate the site of phosphorylation. Underlined entries indicate the residue pair demonstrating statistically significant deviation from independence.
[b]Raw, uncorrected *p*-value is reported when significance is indicated by comparison with empirical multiple hypothesis testing control (see §4).
[c]These amino acid pairs were indicated only by the method of Benjamini & Hochberg [17] and not by the empirical heuristic (see §4*d*).

40 amino acid pairs, enabling the effective identification of lower degrees of energetic cooperativity.

The same methodology was applied to substrates of Cdk1/Cyclin B [13] and CK2 [14]. Strikingly, the set of 71 proline-directed substrates examined of Cdk1/Cyclin B has no pairs of individual amino acids that deviate from a frequency that might be expected by chance at any pair of positions, and among the 432 curated substrates of CK2, only a single amino acid pair at a single pair of positions deviates significantly from what might be expected by chance (table 1). We reasoned that, as in the case of ATM and ATR, we might be able to boost the power of our significance testing procedure by engineering a situation where fewer than the normal 400 pairs of amino acids were accessible. To facilitate this, we combined the 20 amino acids into six functional categories: acidic, basic, hydrophobic, polar, aromatic and a final 'structural' group for the amino acids proline and glycine. Even after the space of possible amino acid pairs is reduced approximately 11-fold in this manner, from 20×20 to 6×6, only a surprisingly small number of statistically significant deviations be found: a single instance for Cdk1/Cyclin B, and four more for CK2.

Despite the use of some of the cleanest and broadest data available on the *in vivo* substrates of protein kinases, there is only in rare instances statistically compelling evidence for interpositional dependencies.

(b) *Utility of interpositional dependencies in predicting novel kinase substrates*

Although the number of statistically significant deviations from a position-independent description of substrate specificity for these kinases is strikingly small, this lack of significant findings is not, of itself, evidence that the kinase recognize their substrates in a position-independent manner. The smaller the cooperative (or anti-cooperative) energetic effect of a pair of amino acids on kinase binding or catalysis, the larger the sample size of substrate sequences that would be required to reliably detect it. The fact that we find only a small number of interdependent sequence elements indicates either that these kinases largely recognize their substrates in a position-independent manner, or that the predominant level of cooperativity is too small to be detected given the available numbers of substrate sequences.

We chose, therefore, to take a step back and examine the issue of substrate sequence interdependencies at a more pragmatic level. The most common motive in trying to determine a substrate specificity descriptor for a kinase is to predict new substrates and new phosphorylation sites. We reasoned that, though very few amino acid pairs at pairs of positions among kinase substrates exhibit statistically significant interdependence, there might be sub-significant interdependencies contributing in aggregate to kinase substrate specificity. To assess directly whether the inclusion of second-order information improved models of kinase substrate specificity, first- and second-order probabilistic models of the specificities of ATM/ATR, CDK1/Cyclin B and CK2 were built and compared (see §4*f*,*g*).

For each kinase, 1000 first- and second-order models were trained on a randomly selected training set of 90 per cent of the substrates, and then tested for their ability to identify the withheld true substrate test set from among a background of shuffled negative control substrates (figures 1*a* and 2*a*). These shuffled controls are generated by shuffling amino acids among substrates, while maintaining their positions relative to the site of phosphorylation. This procedure preserves the position-wise amino acid frequencies of the substrates, but randomizes pairwise interdependencies. Neither a first-order nor a second-order model does a particularly good job of identifying true substrates from among these shuffled negative controls. The first-order models score potential substrates of the modelled kinase having a given sequence as a function of the probabilities of individual amino acids appearing in individual positions among the training data, neglecting all higher-order effects. The second-order models score substrates

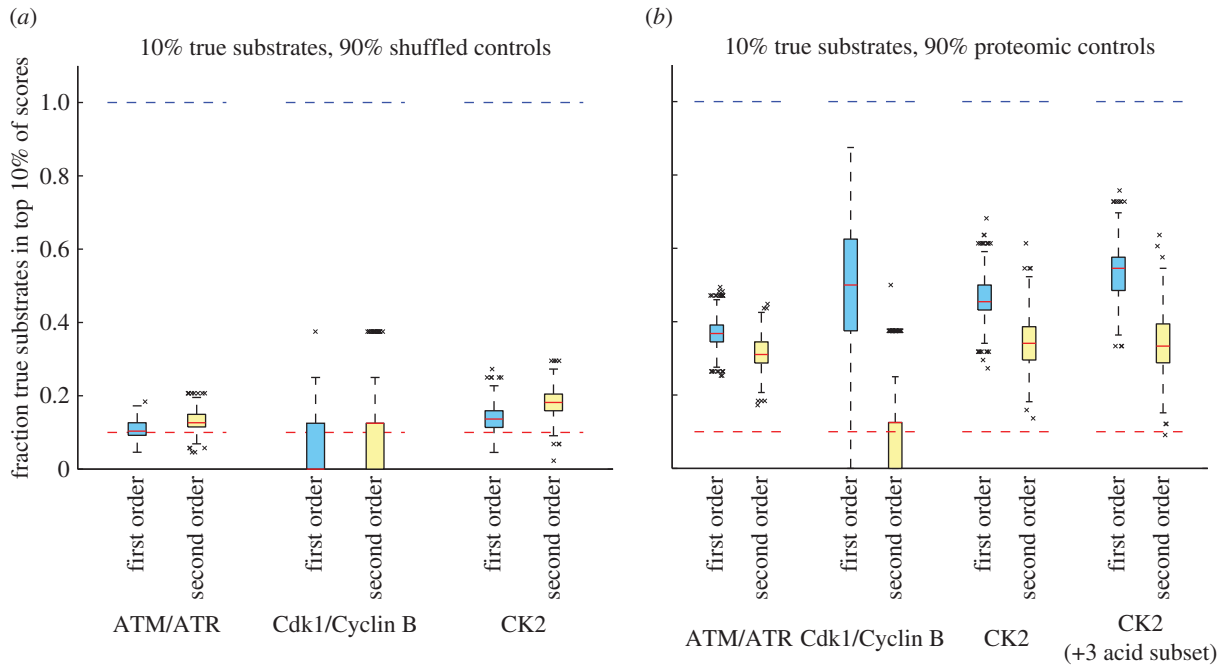

Figure 1. A comparison of the ability of first- and second-order models to correctly identify true kinase substrates. A field of true kinase substrates withheld from training was hidden among mock substrates for each kinase. A field of 10% true and 90% mock substrates was scored using first- and second-order models, and the fraction of true substrates in the top 10% of highest-scoring sequences was counted. The procedure was repeated 1000 times. Plotted boxes span the 25th to 75th percentile of values, with the red line in the boxes marking the medians. Whiskers extend 1.5 times the distance between the 25th and 75th percentiles, and any points more distant from the median are explicitly plotted. Red- and blue-dashed lines at the values 10% and 100% represent the fraction of true substrates expected in the top 10% of scores expected if true and mock substrates were scored randomly, and the maximum possible fraction of true substrates in the top 10% of scores, respectively. (*a*) Mock substrates generated by shuffling true substrates to maintain the probability of each amino acid at each position while breaking interpositional dependencies. (*b*) Mock substrates chosen by randomly selecting sequences from the human proteome conforming to basic known elements of kinase specificity: 'pS/pT-P' for ATM/ATR, 'pS/pT-P' for CDK1/Cyclin B and 'pS/pT-X-X-D/E' for CK2. Because CK2 phosphorylates a number of true substrates that do not have +3 D/E, the CK2 models were trained and tested both with all substrate sequences and with only +3 D/E sequences included.

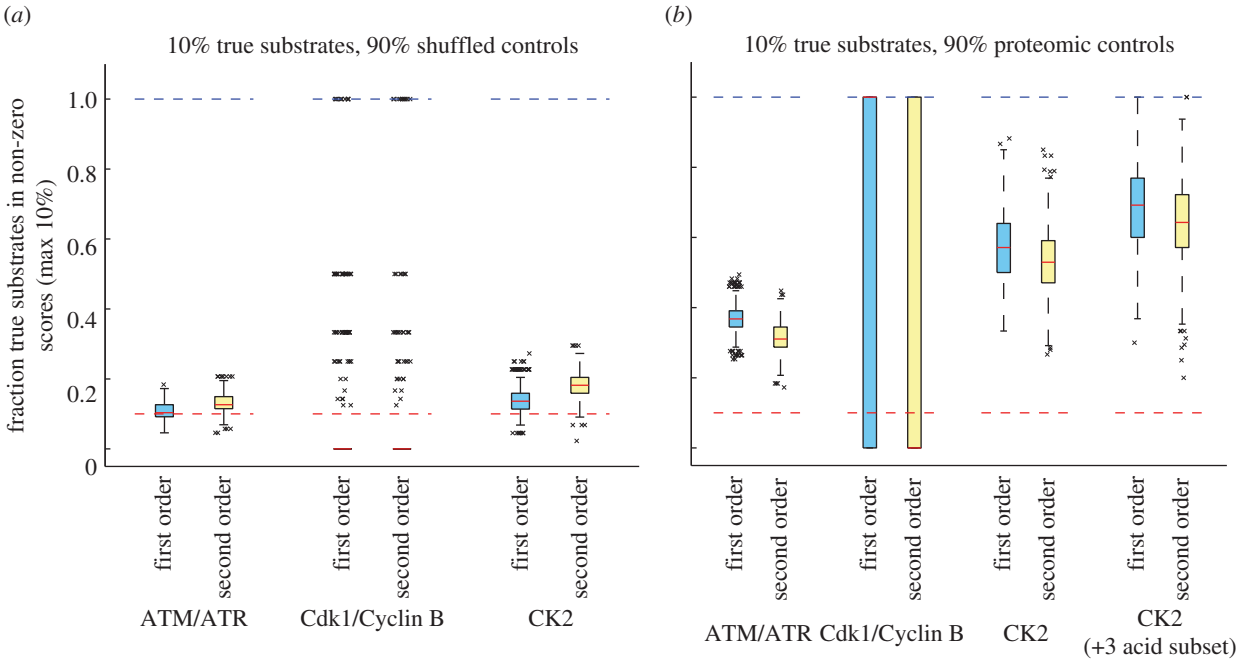

Figure 2. A comparison of the ability of first- and second-order models to correctly identify true kinase substrates, correcting for occurrence of amino acid pairs not present among training data. True kinase substrates were predicted as in figure 1, but rather than examining the top 10% of scored test sequences, a number of sequences for each random splitting of test and training data was examined equal to the least of: 10% of the tested sequences, or the number of tested sequences given a non-zero score under the first- or second-order model. (*a*) Mock substrates generated by position-wise shuffling of true substrates. (*b*) Mock substrates chosen by randomly selecting sequences from the human proteome conforming to basic known elements of kinase specificity.

accounting for both the individual probabilities at single positions and the probabilities of pairs of amino acids at pairs of positions, neglecting the contribution of triplet and higher-order combinations (see §4*f*). These models are used to score a set of 870 potential ATM/ATR substrates (87 true substrates, 783 shuffled negative control substrates). The top 10 per cent (87/870) of scores are taken as predicted substrates from the models. First-order models captured a median value of 10.3 per cent (9/87) true positives among putative hits, whereas the second-order models capture a median value of 12.6 per cent (11/87) true positives (figure 1*a*). Similar results are seen for CK2 (first-order, 13.6% (6/44); second-order, 18.2% (8/44)). Random selection of sequences, as an important point of comparison, would produce 10 per cent true positives.

When the same procedure was applied to CDK1/Cyclin B, we observed that less than 10 per cent of the test data were scored with a non-zero probability of being a kinase substrate according to the second-order model, resulting in supplementation of predicted substrate sites with a random selection of positive and mock test substrates to make up a total of 10 per cent of the test data selected (see figure 1*a* for results, electronic supplementary material, figure S1 for receiver operating characteristic (ROC) curves). To combat this, we also examined the fraction of true positives, restricting the number of predictions analysed for each of 1000 random partitions of training and test data to the least of three quantities: 10 per cent of the test data, or the number of non-zero probability predictions in the first- or second-order model (figure 2*a*). Over 1000 such partitions, a mean of 1.32 sequences per test dataset are assigned a non-zero score. For both the first- and second-order models, the median fraction of these that are true positives is 0 per cent. In each case, then, the second-order model fails to perform compellingly better than the first-order model, and in fact, all models are similar in quality to random selection of peptides from among the test and mock data. The small improvement of both the first- and second-order models over random selection implies that the shuffled negative control substrates are, at the level of sequence, similar in quality to the true substrates.

Identifying true substrates from among mock substrates with an identical position-specific amino acid distribution is, clearly, a difficult task. We reasoned that a less stringent, and more realistic and pragmatic test would be the identification of true substrates from among a background of potential proteomic substrates (figures 1*b* and 2*b*). Such a background can be generated by randomly selecting from the proteome peptides that conform to the most conserved elements of the substrate specificity of the kinase. For ATM/ATR, the motif 'S/T-Q' was used to select mock substrates. For CDK1/Cyclin B, the motif 'S/T-P' was used. In both cases, the selected motif is almost invariably a feature of true substrates. In the case of CK2, the motif 'S/T-X-X-D/E' is common, occurring in about 75 per cent of the substrates in the dataset used. Proteomic peptides conforming to this motif were randomly selected and used to test CK2 substrate specificity models created using either the full dataset, or just the subset that itself conforms to the motif. Again, 90 per cent of the substrate data were used to train models, and then the other 10 per cent of true substrates and a nine-to-one excess of proteomically derived mock substrates were scored, and the top 10 per cent of high-scoring peptides were taken as positively scored (figure 1*b*). The process was repeated 1000 times, and the median number of true positives was noted. Because less than 10 per cent of peptides tested for CDK1/cyclin B and CK2 had non-zero probability scores, the procedure was repeated, limiting the number of examined top-scoring substrates to 10 per cent, or the fewer of non-zero-scoring peptides in the first- or second-order model (figure 2*b*).

In this context of proteomically derived mock substrates, all models had a much better predictive capacity than that observed when attempting to identify true kinase substrates from among negative control substrates generated by position-wise shuffling of the true substrate sequences. When first-order models of ATM/ATR specificity were applied to a field of 10 per cent true substrates and 90 per cent proteomically derived mock substrates, a median of 36.8 per cent (32/87) of the top 10 per cent of high-scoring sequences were true positives, versus 31.0 per cent (27/87) for second-order models (figure 1*b*). Drastically, for CDK1/Cyclin B, 50.0 per cent (4/8) of hits were true positives for first-order models, whereas second-order models captured only 12.5 per cent (1/8) true positives. Models of CK2 substrate specificity were similarly effective at separating true substrates from 'S/T-X-X-D/E' sequences in the proteome, whether trained and tested using the entire CK2 dataset (first-order, 45.5% (20/44); second-order, 34.1% (15/44)) or only the portion itself conforming to the 'S/T-X-X-D/E' motif (first-order, 54.6% (18/33); second-order, 33.3% (11/33)). As was true for CDK1/Cyclin B when tested using position-wise-shuffled negative controls, less than 10 per cent of peptides tested for CDK1/Cyclin B and CK2 had non-zero probability scores. The procedure was therefore repeated limiting the number of examined top-scoring substrates to 10 per cent, or the fewer of non-zero-scoring peptides in the first- or second-order model (figure 2*b*). Very few test substrates for CDK1/Cyclin B, whether true substrates or proteomic mock substrates, receive non-zero scores—none for 883 of 1000 trials, and an average of 1.04 among the remaining 117 trials. While the median fraction of these peptides that are true positives is 100 per cent in the first-order case and 0 per cent in the second-order case among the 117 trials, this is largely an artefact of the fact that most of these 117 trials have only one sequence with non-zero score, making the test largely binary in nature. The means of the same distributions are 71.4 per cent and 41.5 per cent for the first- and second-order models, respectively.

While all models do better than random selection at separating true from proteomically derived mock substrates, in each case, the second-order model is a less accurate predictor than the first-order model, indicating that the pairwise amino acid frequencies among pairs of positions in the training data are not representative of the pairwise amino acid frequencies among the test data. The second-order models are thus systematically overfit to the training data.

3. DISCUSSION

Although the occurrence of significant cooperativity (or anti-cooperativity) is rare among the ATM/ATR, CDK1/Cyclin B and CK2 substrates examined here, a relatively small number of pairs of amino acids that act cooperatively in the bulk context were identified. Most of these do not lend themselves to facile biophysical explanation, with the possible exception of the statistically significant preference for serine over threonine as the phosphorylatable residue in the context of ATM/ATR substrates with proline or glycine in the −1 position, glycine in the +2 position or serine in the +3 position. The difference between a serine and a threonine side chain is quite modest: the threonine has an additional methyl group attached to the beta carbon. The disfavouring of threonine at the phosphorylated position in concert with specific residues at other positions among ATM and ATR substrates is likely caused by one of two biophysical effects: either a steric clash between the threonine methyl group and the residue disfavoured in tandem, or the threonine methyl disfavouring a substrate backbone configuration amenable to kinase–substrate interaction in the context of the other residue.

While crystal structures of the kinases studied here in complex with substrates do not exist, there is a crystal structure solved of the kinase Cdk2/Cyclin A in complex with an optimized substrate peptide [18]. Cdk2 shares 66 per cent sequence identity with Cdk1, and has similar substrate specificity as well. In this structure, there are no close contacts between the amino acid side chains of the substrate peptide, perhaps partially explaining the lack of interdependence seen among Cdk1/Cyclin B substrates.

None of the datasets used in this work is perfectly suited to the task of describing the bulk substrate specificity of a kinase. Although ATM and ATR have a large number of substrates determined in the course of a single study, the two are individual kinases treated, perhaps incorrectly [19], as having identical specificity. Moreover, substrates of the kinases were identified using a cocktail of antibodies [12], and the specificities of these antibodies must be convoluted with those of ATM and ATR to generate the final list of putative substrates. Nonetheless, if ATM preferred different amino acids at individual positions than ATR, then pairs of residues independently preferred at pairs of positions by each of the two kinases would have appeared as being enriched relative to what would be expected under a position-independent model. The curation of CK2 substrates [14], no matter how expertly performed, is subject to the study biases of those who originally reported CK2 substrates in the literature. As with ATM/ATR, however, these biases would be expected to introduce, and not negate, apparent interpositional dependencies. CDK1/Cyclin B substrates were identified by *in vitro* phosphorylation of lysates with an engineered kinase [13], but the number of substrates identified was small with respect to what might be needed to adequately describe the frequency of pairs of amino acids at pairs of positions. Strikingly, the three data sources examined span a wide range of sizes and cover several collection methodologies. Across all three cases, the same consistent pattern of rare interpositional interaction is found.

The methodology applied here is fundamentally similar to both the statistical coupling analysis developed in the research group of Ranganathan [20] and to the mutual information method of Gfeller *et al.* [21]. While each of these methods is aimed at identifying interpositional correlations, we chose our method to directly examine statistical significance of enrichment or diminishment of co-occurrence of pairs of amino acids, rather than using information or entropy as an intermediate metric of statistical significance. Our interest is primarily in identifying co-occurrence rates that are poorly explained by chance, and our methods are meant to approach this goal as directly as possible.

The surprising scarcity of amino acid pairs occurring significantly more or less frequently among a kinase's substrates than would be expected if each amino acid were independently recognized by the kinase might have one of, or a combination of, several explanations. First, it may simply be true that kinases largely recognize each amino acid of their substrates independently, although this seems biophysically implausible. Second, it is possible that kinases incubated *in vitro* with an infinitely varied library of potential substrates would express statistically significant preferences. *In vivo*, however, a kinase is exposed to a subset of the possible amino acid sequence space: it is limited in access to peptide sequences encoded by the genome, in the same subcellular localization and structurally accessible. This convolution may lead to the obfuscation of a kinase's pure biophysical preferences. Finally, there is probably a contribution of effect size. It seems unlikely that each substrate subsite is recognized independently in kinases, and rather more likely that the energetic contribution of second-order effects exists but is very modest. The smaller the size of such effect, the larger the sample of substrates necessary to detect it will be.

A low degree of interpositional dependence in kinase substrate specificity has interesting implications for the evolution of phosphorylation sites and of phosphorylation signalling networks. If each substrate sequence position contributes independently to the ability of a kinase to phosphorylate its substrates, then the evolutionary fitness landscape of substrates as a function of the amino acid at each position is smooth, with a single minimum. That is, there are no non-global local minima acting as traps in kinase substrate fitness space—for any non-ideal substrate, there exist one or more single mutations that would improve the fitness of the substrate for the kinase, with no concerted double or higher-multiple mutations necessary.

Work by other research groups demonstrates that this property of position-wise independence is not uniformly common to all components of phosphosignalling pathways. Yaffe *et al.* [22], in studying the phosphopeptide-binding protein 14-3-3, found that substrates adhered to one of two mutually exclusive sequence modes. Liu *et al.* [23] have demonstrated that SH2 domains, which recognize and bind to phosphotyrosine-containing peptides, will vary in their specificity at some substrate positions as a function of what amino acids are present at other positions. For example, the SH2 domain of Crk binds to peptides

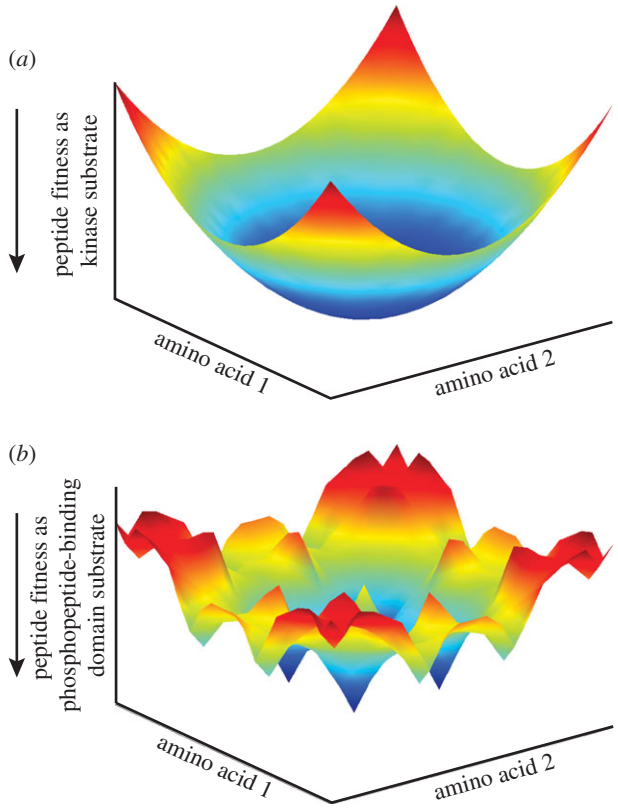

Figure 3. A model of evolutionary fitness landscapes for substrates of kinases and phosphopeptide-binding domains. (*a*) Data presented in this paper indicate that kinase substrate fitness may be position-wise independent in the substrate amino acid sequence, and therefore favourable regions of the substrate fitness space may be accessed relatively easily by a chain of sequential single random mutations. (*b*) Data presented elsewhere [21–23] indicate that phosphopeptide-binding domains express significant interpositional dependencies, indicating that favourable regions of ligand fitness may be separated by energetic barriers.

with a phosphotyrosine along with a leucine or a proline in the +3 position. Proline at the +2 position, however, is allowed only when the identity of the +3 position is leucine and not proline. Other work by Gfeller *et al.* [21] has shown a similar property among the SH3, PDZ and WW peptide-binding domains. Although these domains do not bind phosphopeptides in general, a subset of the WW domains (not explicitly studied by Gfeller *et al.*) does bind specifically to phosphoserine- and phosphothreonine-containing peptides with a proline in the +1 position.

This fundamental difference in the way that kinases and phosphopeptide-binding domains recognize their substrates—kinases in a position-independent manner, but phosphopeptide-binding domains exhibiting clear second- or higher-order preferences—may indicate that the evolution of kinase substrates is a fast or easy step in the evolution of phosphosignalling networks. Single mutations can always improve a non-optimal kinase substrate, whereas the substrates of phosphopeptide-binding domains, which operate in signalling networks to read the phosphorylation events left by kinases, may sometimes require concerted mutations in potential substrates in order to become more fit for binding (figure 3). Evidence for the existence of significant numbers of functionless phosphorylations [24] is consistent with this possibility; a build-up of non-functional phosphorylations is consistent with kinase–substrate evolution being an easy step in signalling network evolution.

It is clear from the results presented here that the specificity of these kinases for the amino acid sequences proximal to the site of phosphorylation among their *in vivo* substrates is largely well-described by a first-order model. Adding second-order information to first-order models of kinase specificity, at least for these kinases, seems to add only a minimal benefit in terms of predicting novel substrates. In some cases, adding second-order information even reduces the quality of a first-order model, indicating that the second-order models are overfit to irrelevant interpositional correlations in the training data. In order to predict novel kinase substrates, it may instead be beneficial to integrate simple sequence models with other contextual information such as known interactions [11], subcellular localization, protein structure and distal-site recognition.

In the present work, we elected to restrict our models to using only local sequence information directly, in order to isolate the effect of second-order information on improving first-order models. Sequence information also informs substrate fitness in indirect ways not examined here. Intrinsic protein disorder seems to be enriched in proximity to sites of protein phosphorylation, and consideration of protein disorder improves *ab initio* prediction of the location of sites of phosphorylation [25]. Likewise, the evolution and conservation of protein sequence surrounding a site of phosphorylation might give clues to which neighbouring residues are responsible for recognition, and what the relevant constraints on flanking sequence are, though aligning sequences in disordered regions of proteins is quite difficult.

It is tempting to speculate that the approximately first-order nature of substrate recognition by these kinases reflects evolutionary freedom for the development of new kinase substrates: the energy landscape for substrate fitness is smooth, with multiple-step mutations generally unnecessary to improve the fitness of any potential kinase substrate. It remains to be seen, however, whether the results reported here extend to other kinases, or to the complete repertoire of substrates for the kinases analysed here.

4. MATERIAL AND METHODS

(a) *Data sources*

We attempted to find long lists of substrates for a variety of kinases, identified in as unbiased a manner as possible. We chose two datasets that identified a long list of substrates via a single experiment: one for ATM/ATR [12] that identified 894 individual human phosphosites with the canonical 'pS/pT-Q' motif of ATM and ATR, and one for Cdk1/Cyclin B [13] that identified 77 human sites, 71 of which contained the canonical proline-directed 'pS/pT-P' motif. Additionally, we identified a third dataset collected by expert curation that enumerates 432 specific sites of CK2 phosphorylation from a variety of species [14].

(b) *Data preparation*

For the ATM/ATR and CDK1/Cyclin B data sources, full protein sequences were retrieved using protein identifier and sequence information from the source publication. In some cases, the source data contained duplications of the same site, or were not sufficient to unambiguously identify the protein and site referenced. The remaining sequences were used to find peptide sequences including the site of phosphorylation, as well as the seven residues upstream and downstream of the phosphorylation. This preparation yielded 861 putative human ATM/ATR sites and 71 putative human Cdk1/Cyclin B sites. Protein identifiers and sequences are provided as electronic supplementary material, tables S1 and S2. The CK2 data were originally published including the sequences of phosphorylation sites, including six residues to the N- and C-terminal sides of the site of phosphorylation itself, and the data were used directly.

(c) *Identification of overrepresented and underrepresented amino acid pairs*

For each kinase, we attempted to identify those pairs of amino acids, at pairs of positions relative to the phosphorylated residue that co-occurred among substrates with a frequency not adequately explained by their individual prevalences and chance. The positions considered were selected by inspection of a motif logo [26] built from substrate sequences as a sequential set of positions that contain more information than the background (see electronic supplementary material, figure S2). For ATM/ATR substrates, positions from -2 to $+3$ relative to the site of phosphorylation were considered. For CDK1/Cyclin B, positions -2 to $+4$ were examined. For CK2, positions -1 to $+5$ relative to the phosphorylation site were considered. For every pair of amino acids at each pair of positions, the statistical significance of deviation from what would be expected by chance, if the positions were recognized independently by the phosphorylating kinase, was calculated using a hypergeometric distribution. If a kinase has N total substrates, with m having amino acid a at position x, n having amino acid b at position y and k having both, then the expression

$$P_E = \sum_{i=k}^{\min(m,n)} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}},$$

gives the statistical significance of enrichment of amino acid a at position x and amino acid b at position y co-occurring. This is exactly equal to the probability of finding k or more sequences with both amino acids at the two positions by chance, assuming the two are not recognized cooperatively. Conversely, the expression

$$P_R = \sum_{i=\max(0,n+m-N)}^{k} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}},$$

gives the significance of reduction, or the probability of finding k or fewer instances of both amino acids at the two positions by chance if they are recognized independently.

(d) *Statistical significance cut-off determination*

Two methods were used to set a cutoff for statistical significance values for the interactions between amino acid pairs that corresponds to a 5 per cent false-positive rate. The first was entirely empirical: 1000 times for each dataset, we randomly shuffled the data at each amino acid position across all sequences, such that the probability of each amino acid at each position is preserved, but any pairwise interactions were scrambled. For each shuffling, the significances of enrichment and reduction of all pairs of amino acids at all pairs of positions were calculated as for the true data. The single most significant result for each of these thousand negative controls was identified. In the true data, any result more significant than can be found in 95 per cent of shuffled controls was taken as significant and presented—that is, we accepted as a positive result any p-value that did not appear in randomized data more than 5 per cent of the time.

The more rigorous false discovery rate control method described by Benjamini & Hochberg [17] was also used, independently on one vector each of statistical significances of enrichment and reduction for each pair of substrate positions for each kinase examined. A q^* value of 0.05 was used, and the set of hypotheses tested were only those that could not be trivially accepted: for enrichment, only those pairs of amino acids that occur at least once apiece independently, and for reduction only those pairs of amino acids that co-occur once or more.

Although the false discovery rate control procedure is only well-suited to series of significance tests that are statistically independent, and the interdependence of statistical significance values calculated by studying the frequency of pairs of amino acids at each pair of positions is difficult to accurately characterize, results were in strong agreement with those generated using the empirical procedure.

(e) *Simplified amino acid alphabet*

Because the numbers of substrates identified for each kinase of interest is small compared with the number of possible pairs of amino acids at each position, all calculations were repeated using a simplified amino acid alphabet comprised of six classes: the acids (D,E), the bases (K,R), the hydrophobes (A,I,L,M,V), the aromatics (F,W,Y), the polar side chains (C,H,N,Q,S,T) and the structural amino acids (G,P). Significant results generated using this simplified alphabet are reported only when no pair of individual amino acids from the pair of classes is itself capable of recapitulating the significant result.

(f) *First- and second-order models of kinase substrate specificity*

Probabilistic models of kinase substrate specificity were generated that approximate the many-order probability distribution across all amino acid positions using only single positions, or using single positions

and pairs of positions, according to the generalized Kirkwood superposition approximation as described by Killian *et al.* [27]. The expression:

$$p^{(1)}(x_{1,\ldots,}x_m) = \prod_{i=1}^{m} p_1(x_i),$$

gives $p^{(1)}$, a first-order approximation of the probability of a given sequence of amino acids x_1 to x_m at positions 1 to m as the product of the probability of each independently. This approximation neglects any cooperative effect of pairwise or higher-order combinations of amino acids. The expression

$$p^{(2)}(x_{1,\ldots,}x_m) = \frac{\prod_{C_2^m} p_2}{\left[\prod_{C_1^m} p_1\right]^{m-2}},$$

gives $p^{(2)}$, a second-order approximation of the same probability that accounts for individual and pairwise influences, but neglects any higher-order effects. The probabilities multiplied in the numerator are all of the probabilities of pairs of amino acids at all pairs of positions, whereas the probabilities in the denominator are those for individual amino acids at single positions as in the first-order approximation.

(g) *Evaluation of kinase substrate specificity models*

Models were trained for each kinase by randomly separating the data to 90 per cent for training to populate probabilistic models as described earlier, and 10 per cent for testing. The testing portion of the data was combined with nine times as many mock substrate sequences, generated in one of two ways. First, mock sequences were generated by shuffling amino acids among the test substrate sequences within each amino acid position, preserving exactly the probabilities of amino acids at each position while randomizing pairwise interactions. Second, mock substrate sequences were generated by selecting appropriate sequences randomly from the human proteome (International Protein Index [28], v3.55). These sequences were selected from among all sites containing the amino acid sequence 'S/T-Q' for ATM/ATR, the sequence 'S/T-P' for CDK1/Cyclin B and the sequence 'S/T-X-X-D/E' for CK2. Because an acid at the +3 position relative to the site of phosphorylation is not an absolute requirement for CK2, additional models were trained and tested against proteomic mock substrates for CK2 using only the subset of 323 substrates conforming to the 'S/T-X-X-D/E' motif. The ability of each probabilistic model to identify true substrates among a background of mock substrates was measured by counting the number of true substrates among the best 10 per cent of potential substrates scored. This procedure was repeated using 1000 random divisions of the substrate data into test and training sets, and the mean and standard deviation are reported. In some cases, the best 10 per cent of potential substrates under a second-order model included some substrates with a probability score of 0. In these cases, the potential substrates with a score of 0 included in the best 10 per cent were selected randomly from all potential substrates with a score of 0.

Because peptides with a 0 per cent probability score would probably not be taken as hits by a researcher using these models to predict substrates, we repeated the procedure taking as a maximum the best 10 per cent of potential substrates, but limiting ourselves to those with a probability score of more than 0. For each of the 1000 random divisions of test and training data, the same number of high-ranking sequences was examined for both the first- and second-order model: the least of 10 per cent of the sequences, the number given a non-zero score by the first-order model, or the number given a non-zero score by the second-order model. If no sequences passed these criteria for one of the 1000 divisions, then this division was not included in distributions to calculate median statistics. This procedure led us to look at 87 of 870 sequences in all 1000 test sets for ATM/ATR, a mean of 2.01 of 80 sequences in 654 of 1000 test sets for CDK1/Cyclin B, and 44 of 440 sequences in all 1000 CK2 test sets, when comparing with position-wise-shuffled substrate sequences. When comparing with proteomic mock substrate sequences, we used 87 of 870 sequences in all 1000 ATM/ATR test sets, a mean of 1.04 of 80 sequences in 117 of 1000 CDK1/Cyclin B test sets and a mean of 25.2 or 15.4 in all 1000 test sets for CK2, when the training and test data did or did not include sequences not matching the 'S/T-X-X-D/E' motif, respectively.

The authors thank members of their research groups, and particularly Dr David Clarke for helpful discussion. This work was supported by NIH grants no. CA112967 (M.B.Y. and D.A.L.), ES015339 (M.B.Y.) and R24-DK090963 (D.A.L.), and by a Singapore-MIT Alliance Fellowship to C.L.

REFERENCES

1 Amanchy, R., Periaswamy, B., Mathivanan, S., Reddy, R., Tattikota, S. G. & Pandey, A. 2007 A curated compendium of phosphorylation motifs. *Nat. Biotechnol.* **25**, 285–286. (doi:10.1038/nbt0307-285)
2 Alexander, J. *et al.* 2011 Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. *Sci. Signal.* **4**, ra42. (doi:10.1126/scisignal.2001796)
3 Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. 2003 Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31**, 3635–3741. (doi:10.1093/nar/gkg584)
4 Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S. & Brunak, S. 2004 Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* **4**, 1633–1649. (doi:10.1002/pmic.200300771)
5 Miller, M. L. *et al.* 2008 Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal.* **1**, ra2. (doi:10.1126/scisignal.1159433)
6 Huang, H. D., Lee, T. Y., Tzeng, S. W. & Horng, J. T. 2005 KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.* **33**, W226–W229. (doi:10.1093/nar/gki471)
7 Kim, J. H., Lee, J., Oh, B., Kimm, K. & Koh, I. 2004 Prediction of phosphorylation sites using SVMs. *Bioinformatics* **20**, 3179–3184. (doi:10.1093/bioinformatics/bth382)

8 Xue, Y., Li, A., Wang, L., Feng, H. & Yao, X. 2006 PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* **7**, 163. (doi:10.1186/1471-2105-7-163)

9 Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J. & Diella, F. 2011 Phospho.ELM: a database of phosphorylation sites: update 2011. *Nucleic Acids Res.* **39**, D261–D267. (doi:10.1093/nar/gkq1104)

10 Farriol-Mathis, N., Garavelli, J. S., Boeckmann, B., Duvaud, S., Gasteiger, E., Gateau, A., Veuthey, A. L. & Bairoch, A. 2004 Annotation of post-translational modifications in the Swiss–Prot knowledge base. *Proteomics* **4**, 1537–1550. (doi:10.1002/pmic.200300764)

11 Linding, R. *et al.* 2007 Systematic discovery of *in vivo* phosphorylation networks. *Cell* **129**, 1415–1426. (doi:10.1016/j.cell.2007.05.052)

12 Matsuoka, S. *et al.* 2007 ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science* **316**, 1160–1166. (doi:10.1126/science.1140321)

13 Blethrow, J. D., Glavy, J. S., Morgan, D. O. & Shokat, K. M. 2008 Covalent capture of kinase-specific phosphopeptides reveals Cdk1-cyclin B substrates. *Proc. Natl Acad. Sci. USA* **105**, 1442–1447. (doi:10.1073/pnas.0708966105)

14 Salvi, M., Sarno, S., Cesaro, L., Nakamura, H. & Pinna, L. A. 2009 Extraordinary pleiotropy of protein kinase CK2 revealed by weblogo phosphoproteome analysis. *Biochim. Biophys. Acta* **1793**, 847–859. (doi:10.1016/j.bbamcr.2009.01.013)

15 Meggio, F. & Pinna, L. A. 2003 One-thousand-and-one substrates of protein kinase CK2? *FASEB J.* **17**, 349–368. (doi:10.1096/fj.02-0473rev)

16 Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E. & Zhang, B. 2004 PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4**, 1551–1561. (doi:10.1002/pmic.200300772)

17 Hochberg, Y. & Benjamini, Y. 1990 More powerful procedures for multiple significance testing. *Stat. Med.* **9**, 811–818. (doi:10.1002/sim.4780090710)

18 Brown, N. R., Noble, M. E., Endicott, J. A. & Johnson, L. N. 1999 The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. *Nat. Cell Biol.* **1**, 438–443. (doi:10.1038/15674)

19 Kim, S. T., Lim, D. S., Canman, C. E. & Kastan, M. B. 1999 Substrate specificities and identification of putative substrates of ATM kinase family members. *J. Biol. Chem.* **274**, 37 538–37 543.

20 Lockless, S. W. & Ranganathan, R. 1990 Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299. (doi:10.1126/science.286.5438.295)

21 Gfeller, D. *et al.* 2011 The multiple-specificity landscape of modular peptide recognition domains. *Mol. Syst. Biol.* **7**, 484. (doi:10.1038/msb.2011.18)

22 Yaffe, M. B., Rittinger, K., Volinia, S., Caron, P. R., Aitken, A., Leffers, H., Gamblin, S. J., Smerdon, S. J. & Cantley, L. C. 1997 The structural basis for 14-3-3: phosphopeptide binding specificity. *Cell* **91**, 961–971. (doi:10.1016/S0092-8674(00)80487-0)

23 Liu, B. A., Jablonowski, K., Shah, E. E., Engelmann, B. W., Jones, R. B. & Nash, P. D. 2010 SH2 domains recognize contextual peptide sequence information to determine selectivity. *Mol. Cell. Proteomics* **9**, 2391–2404. (doi:10.1074/mcp.M110.001586)

24 Landry, C. R., Levy, E. D. & Michnick, S. W. 2009 Weak functional constraints on phosphoproteomes. *Trends Genet.* **25**, 193–197. (doi:10.1016/j.tig.2009.03.003)

25 Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z. & Dunker, A. K. 2004 The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32**, 1037–1049. (doi:10.1093/nar/gkh253)

26 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. 2004 WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190. (doi:10.1101/gr.849004)

27 Killian, B. J., Yundenfreund Kravitz, J. & Gilson, M. K. 2007 Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.* **127**, 024107. (doi:10.1063/1.2746329)

28 Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E. & Apweiler, R. 2004 The International Protein Index: an integrated database for proteomics experiments. *Proteomics* **4**, 1985–1988. (doi:10.1002/pmic.200300721)